STEWARD: Demo of Spatio-Textual Extraction on the Web Aiding the Retrieval of Documents[*]

Hanan Samet
Michael D. Lieberman
Jagan Sankaranarayanan
Department of Computer Science
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland at College Park
{hjs,codepoet,jagan}@cs.umd.edu

Jon Sperling
HUD Office of Policy Development & Research (PD&R)
451 7th St SW, Rm 8146
Washington D.C. 20410
jon_sperling@hud.gov

ABSTRACT

A spatio-textual search engine, termed "STEWARD" is demonstrated where document similarity is based on both the textual similarity as well as the spatial proximity of the locations in the document to the spatial search input. STEWARD's performance is enhanced by the presence of a document tagger that is able to identify textual references to geographical entities. The user-interface of STEWARD provides the ability to browse results, thereby making it a valuable "knowledge discovery" tool.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms,Design,Performance

Keywords

Spatio-textual search engine, STEWARD, Geocoding

1. INTRODUCTION

Search technology today is dominated by search engines such as the one provided by Google where documents are retrieved with the aid of an algorithm that ranks documents related to the query string on the basis of how many other documents link to it. We are interested in developing a search engine where the query string contains a geographical entity and we wish to find other documents that are related to it by spatial proximity. For example, a document containing "Los Angeles" is deemed relevant to a query string containing "Hollywood", even though the query string "Hollywood" might not even be mentioned in the document. In this paper we describe a demo of STEWARD (denoting "Spatio-Textual Extraction on the Web Aiding the Retrieval of Documents"), a spatio-textual search engine for retrieval of documents on the HUDUSER.ORG web site under development at the University of Maryland in cooperation with the Office of Policy Development & Research of the Department of Housing and Urban Development (HUD PD&R).

2. STEWARD

Queries to STEWARD can have a purely geographical component, a keyword component, or a combination of both. When the query string is purely a geographical entity, we wish to find documents that are related to it by spatial proximity. The documents that are returned are ranked by the extent to which STEWARD determines that the geographic entity that forms the query string serves as the geographic focus of the document. This is based on many factors which include the number of times that the search string or proximate geographic locations are mentioned in the document.

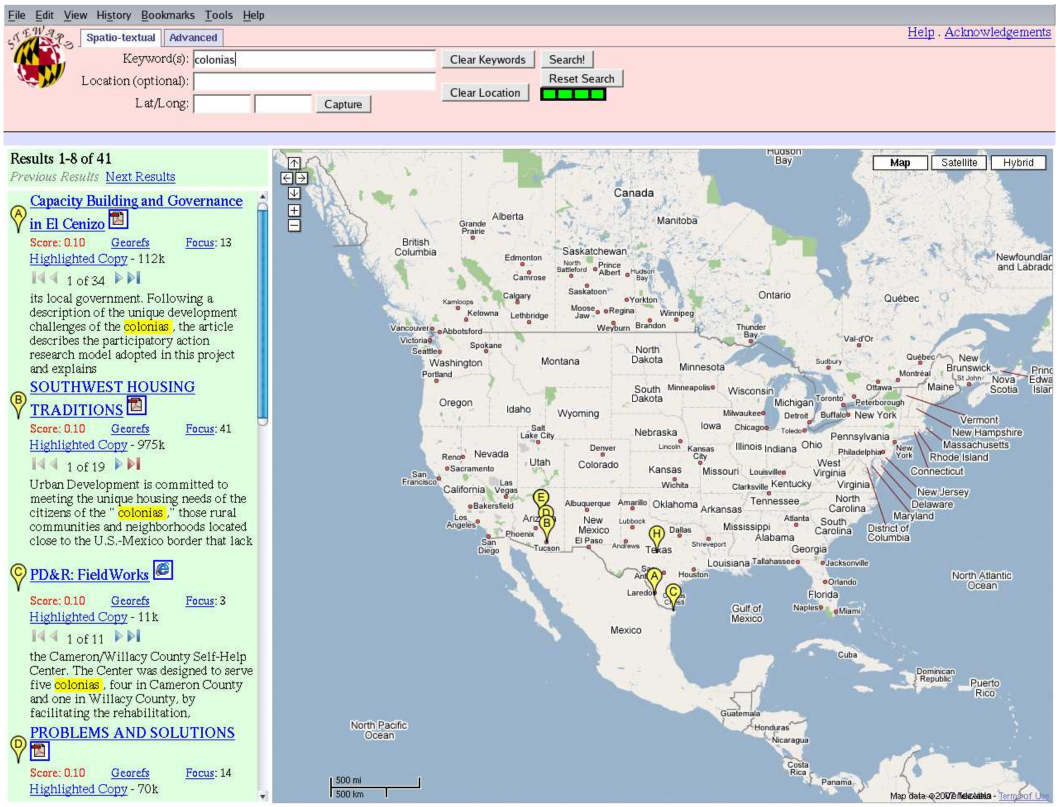
STEWARD's notion of a geographic focus differs from much of the existing work in this area which has been cast in terms of finding the geographic scope of web sites which contain one or more documents and is usually done by examining their link structure. Instead, our focus is on the actual contents of documents. Moreover, we are not only interested in finding a geographic focus sufficiently general to span the entire document, but instead also wish to identify as many geographical locations as possible as well as provide the ability to browse through the documents in order of spatial proximity to the designated keywords.

STEWARD uses a document tagger built by us to identify potential references to geographic locations in unstructured text documents. The tagger is aware of sentence structure so that proper nouns can be determined. The determination of which of these nouns or word combinations are indeed geographic locations is facilitated with the aid of a gazetteer such as that available in GNIS for the United States and its analog GEONET for foreign names. Of course, there is still the issue of distinguishing between multiple locations with the same name such as, for example, "Springfield, IL" and "Springfield, MA", which is non-trivial.

When the queries consist only of a non-geographic keyword, then STEWARD ranks the documents on the basis of the strength of the occurrences of the keyword in them. In addition, in this case, STEWARD also identifies all of the references to geographic locations in each document and ranks them in the order in which it determines that they serve as the geographic focus of the document. This is based, in part, on the frequency of their occurrence, as well as the distribution of their occurrences in the documents.

When a geographic location is presented as input to STEWARD along with input keywords, the relevant documents (i.e., the ones containing an instance of the input keywords) are ranked in in-

[*]This work was supported in part by the US National Science Foundation under Grant EIA-00-91474 and CCF-05-15241, as well as the Office of Policy Development & Research of the Department of Housing and Development (HUD PD&R)



Keyword(s): colonias

Location (optional):

Lat/Long:

Results 1-8 of 41

Previous Results Next Results

Ⓐ Capacity Building and Governance in El Cenizo

Score: 0.10 Georefs Focus: 13

Highlighted Copy - 112k

1 of 34

its local government. Following a description of the unique development challenges of the colonias , the article describes the participatory action research model adopted in this project and explains

Ⓑ SOUTHWEST HOUSING TRADITIONS

Score: 0.10 Georefs Focus: 41

Highlighted Copy - 975k

1 of 19

Urban Development is committed to meeting the unique housing needs of the citizens of the " colonias ," those rural communities and neighborhoods located close to the U.S.-Mexico border that lack

Ⓒ PD&R: FieldWorks

Score: 0.10 Georefs Focus: 3

Highlighted Copy - 11k

1 of 11

the Cameron/Willacy County Self-Help Center. The Center was designed to serve five colonias , four in Cameron County and one in Willacy County, by facilitating the rehabilitation,

Ⓓ PROBLEMS AND SOLUTIONS

Score: 0.10 Georefs Focus: 14

Highlighted Copy - 70k

Figure 1: Example screenshot of the STEWARD system.

creasing order of distance of their geographic focus from the geographic location component of the query string. The geographic location component of the input query can be expressed in terms of latitude/longitude or as a textual reference to a spatial object. For example, the user could search for “Housing Projects” in the vicinity of “College Park, MD”. The results would only return such documents that qualify both the content and location specifier that was provided to the system by the user.

A preliminary version of the STEWARD system is currently being deployed on the HUDUSER.ORG web site and is available to anyone with an Internet connection through an interactive user interface written in HTML and AJAX. Figure 1 shows a screenshot of the user-interface of the STEWARD system running on the Mozilla Firefox browser in response to a purely textual query seeking all documents containing the keyword “colonias” which are settlements lying primarily along the US-Mexico border.

From the figure we see that the user-interface is divided into three panes (i.e., regions). The top pane is being used to specify the query parameters via text boxes for the textual keyword as well as a location, which is optional. The left pane shows the documents that satisfy the textual keywords, along with a small extract showing the context in which the keyword is found. The right pane positions the documents that satisfy the textual keyword in the query on the map using icons at positions that STEWARD has determined to be their geographic focus. In this case, we find that the geographic foci of these documents do indeed lie on the US-Mexico border, which is not surprising, but it is reassuring that STEWARD has correctly identified them. In addition, the right pane can be used to input the desired location for the geographic scope of a query. Documents that satisfy the textual keyword are reported in increasing order of the distance of their geographic focus from the query point. Notice the clean separation of the textual results from the spatial results in the user interface.

STEWARD also enables users to browse through the relevant documents that it has found, and to highlight, in sequence, all occurrences of the keyword. In addition, each of the relevant documents can be browsed to show all occurrences of each of the geographic locations that it has found in the document, or to show the most important occurrence of each of these geographic locations. In the former, STEWARD simply provides an extract of the context in which the location appears, while for the latter, a pointer to the geographic location is also provided along with the extract.

3. CONCLUDING REMARKS

Many enhancements are planned to increase the power of STEWARD as well as its user interface. In particular, the spatial querying capability of STEWARD will be augmented to include capabilities present in the SAND Spatial Browser [1]. This also includes adding the ability to draw the extent of the spatial queries on the map rather than being restricted to a textual specification. Moreover, a mechanism for users to provide feedback on the quality of the search will be added.

4. REFERENCES

[1] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *Communications of the ACM*, 46(1):63–66, Jan. 2003.